

Contents lists available at [SciVerse ScienceDirect](http://SciVerse.ScienceDirect.com)

## Journal of Biomedical Informatics

journal homepage: [www.elsevier.com/locate/yjbin](http://www.elsevier.com/locate/yjbin)

# Toward automated consumer question answering: Automatically separating consumer questions from professional questions in the healthcare domain

Feifan Liu <sup>a,\*</sup>, Lamont D. Antieau <sup>a</sup>, Hong Yu <sup>a,b</sup><sup>a</sup> Department of Health Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI, United States<sup>b</sup> Department of Computer Science, University of Wisconsin-Milwaukee, Milwaukee, WI, United States

## ARTICLE INFO

## Article history:

Received 7 March 2010

Accepted 5 August 2011

Available online 12 August 2011

## Keywords:

Question classification

Medical question answering

Supervised machine learning

Support vector machines

Natural language processing

## ABSTRACT

**Objective:** Both healthcare professionals and healthcare consumers have information needs that can be met through the use of computers, specifically via medical question answering systems. However, the information needs of both groups are different in terms of literacy levels and technical expertise, and an effective question answering system must be able to account for these differences if it is to formulate the most relevant responses for users from each group. In this paper, we propose that a first step toward answering the queries of different users is automatically classifying questions according to whether they were asked by healthcare professionals or consumers.

**Design:** We obtained two sets of consumer questions (~10,000 questions in total) from Yahoo answers. The professional questions consist of two question collections: 4654 point-of-care questions (denoted as PointCare) obtained from interviews of a group of family doctors following patient visits and 5378 questions from physician practices through professional online services (denoted as OnlinePractice). With more than 20,000 questions combined, we developed supervised machine-learning models for automatic classification between consumer questions and professional questions. To evaluate the robustness of our models, we tested the model that was trained on the Consumer–PointCare dataset on the Consumer–OnlinePractice dataset. We evaluated both linguistic features and statistical features and examined how the characteristics in two different types of professional questions (PointCare vs. OnlinePractice) may affect the classification performance. We explored information gain for feature reduction and the back-off linguistic category features.

**Results:** The 10-fold cross-validation results showed the best F1-measure of 0.936 and 0.946 on Consumer–PointCare and Consumer–OnlinePractice respectively, and the best F1-measure of 0.891 when testing the Consumer–PointCare model on the Consumer–OnlinePractice dataset.

**Conclusion:** Healthcare consumer questions posted at Yahoo online communities can be reliably classified from professional questions posted by point-of-care clinicians and online physicians. The supervised machine-learning models are robust for this task. Our study will significantly benefit further development in automated consumer question answering.

© 2011 Elsevier Inc. All rights reserved.

## 1. Introduction

The use of computer technology is an integral part of meeting healthcare information needs. While the earliest use of computers to search for medical information was generally performed by healthcare professionals, the Internet has paved the way for laypeople with healthcare concerns such as diseases, symptoms, and treatments to search for information [1–3], and the number of people using the Internet for such information is growing. Tu and Cohen [4] report, for instance, that the number of those searching for health-related information online doubled between 2001 and

2007 to include nearly one-third of all adult Internet users. This growth has not only expanded the healthcare topics that users search for, but also expanded in the myriad sources available to them; recently, for instance, there has been a boom in healthcare-related social networking [5], particularly among question-answering sites [6]. These trends suggest that healthcare consumers will become increasingly dependent on Internet resources for answers to their medical questions in the years ahead.

In light of these developments, this paper reports on efforts to improve the medical question-answering system, AskHERMES (<http://www.askhermes.org/>). AskHERMES (Help clinicians Extract and articulate Multimedia information from literature to answer their ad hoc clinical quEstions) is a fully automated system that uses natural language processing tools to retrieve, extract, analyze, and integrate information from medical literature and other

\* Corresponding author. Address: 2400 E Hartford Ave., Room 953, Milwaukee, WI 53211, United States. Fax: +1 414 229 2619.

E-mail address: [liuf@uwm.edu](mailto:liuf@uwm.edu) (F. Liu).

information resources to provide answers in response to questions posed by healthcare professionals, e.g. physicians. To date, much of the research that has been done to benefit the system has focused on extracting information needs from the complex questions that physicians often ask [7] and to present responses to such questions in the most effective way [8]. More recently, the use of automatic speech recognition tools that enable physicians to pose questions to the system as spoken rather than typewritten input has been explored [9]. The research direction presented in the current paper, however, reflects the belief that AskHERMES could also benefit healthcare consumers, but only after the development of several subcomponents will it accurately interpret the information needs of lay users (i.e. healthcare consumers), quickly find information appropriate to their literacy level and technical expertise (or lack thereof), and then summarize the information that is retrieved and present it in the most useful way.

As a first step in this process, we investigated ways of determining whether users are healthcare consumers or professionals based on quantitative linguistic differences between the questions asked by members of both groups. Although the literature on differences in communication between physicians and patients acknowledges that questions are a significant component of medical encounters [10–12], studies in this area have generally focused on questions that patients ask healthcare professionals [10–14], questions posed by healthcare professionals [13], or on the more general aspects of linguistic interaction between the two groups [15]. With respect to the literature on using computers to search for medical-related information, studies have also tended to investigate either the queries of healthcare consumers [16] or those of professionals [17] without taking into account the questions of both groups. A more rigorous, comparative analysis of these questions might reveal stylistic differences that could enable us to better meet the information needs of members of both groups.

With these aims in mind, we developed a supervised machine-learning framework to automatically distinguish the questions of healthcare consumers and professionals. Although an exploratory study of the differences between the information-seeking behaviors of the two groups revealed significant differences at every level of the grammar, we primarily focused on the shallow-level linguistic features (e.g. bag-of-words features) without deep language processing (e.g. syntactic parsing), as previous work determined that words are adequate representational units for the purposes of classification [18]. We found machine-learning approaches suitable for classifying questions by whether the question was posed by consumers or professionals. In addition to the success of a bag-of-words approach for classification, we experimented with statistical features and linguistic category features to improve the robustness of the classifiers.

## 2. Related work

Many studies in question classification have focused on the semantics of questions and their potential answers, and to that end, they have investigated the use of taxonomies in question classification both in the open domain [19,20] and in the medical domain [21]. Some systems have explored the use of syntactic features for classification but have generally done so as a supplement to semantics rather than as a replacement [22–24]. Other studies have identified additional dimensions that could be useful for question classification, for instance, the distinction between factual and analytical questions [25,26], factual and opinion questions [27], objective and subjective questions [28,29], and answerable and unanswerable questions [30]. We propose that the ability to distinguish between the questions asked by consumers and professionals could be a dimension worth exploring, in our case,

because of its potential to tailor information retrieval and question answering systems for different users.

Different linguistic features and feature selection methods have been studied in previous work. In the area of corpus linguistics, studies focusing on readability [31–33] have explored word length, word frequency, and sentence length to determine linguistic complexity and genre. The information gain based feature selection has shown to be helpful for text and evidence classification [34,35]. Motivated by those prior works, we evaluated both linguistic features and statistical features on our task, and the proposed linguistic category features which are expected to capture language usage differences on a higher level between healthcare professionals and consumers, thereby eluding the data sparseness problem resulting from “bag of words” features.

## 3. Material and methods

We first discuss the collection of our data and provide a brief characterization before describing the machine-learning methods that were used for question classification.

### 3.1. Data

We used four representative datasets in our study: two sets of consumer questions and two sets of professional questions, as described below.

#### 1. Consumer questions I (Consumer-I):

We downloaded 5013 consumer questions posted on Yahoo Answers between May and June 2009 (<http://answers.yahoo.com>, category “Health/Diseases and Conditions”).

#### 2. Consumer questions II (Consumer-II):

We reused 5499 consumer questions, which is a subset extracted from a previous study <http://ir.mathcs.emory.edu/shared/>. Questions in this subset were posted in the “Health/Diseases and Conditions” category on Yahoo Answers between Nov. 2007 and Jan. 2008.

#### 3. Point-of-care clinical questions (PointCare):

A set of 4654 professional questions collected by physicians from interviews of family doctors following patient visits [13,36].

#### 4. Online questions among physician practices (OnlinePractice):

ParkHurstExchange (<http://www.parkhurstexchange.com>) is an online publishing service based in Montreal, Canada, which provides credible and highly respected publications of physician practice questions and answers from healthcare professionals. All questions posted by physicians are selected and answered by professional members, which are further reviewed and approved by the Medical Editor-in-Chief. Through this service, physicians can ask their own questions, browse questions in different specialties and search them by keywords. We downloaded 5378 professional questions from the ParkHurstExchange website as of December 6, 2010.

Although all four collections of questions described above were not intended for an automated question-answering system, there are several benefits of using these question collections: (1) they are relatively large collections of questions in which each question can be attributed to a consumer or a professional with a high degree of certainty and thus are amenable to supervised

machine-learning processes; (2) they go beyond the use of search terms to include utterances comprising complete sentences and even longer passages of discourse, which we anticipate will be more representative of queries that are formulated via natural language as opposed to keywords and phrases; and (3) two different professional datasets represents two different clinical settings based on varying factors; for instance, point-of-care questions are relatively spontaneous while online physician practice questions are relatively well-planned. This allows us to examine the diversity effects on the classification performance and robustness of our proposed approach.

### 3.2. Linguistic observation on question collections

There is a wide range of question types in both consumer and professional questions, and typologies for the healthcare professional questions have been proposed in a number of publications [13,36]. A typological classification based on the interrogative words in questions [36] is not only useful for the professional questions, but also applies to the Yahoo consumer questions, as shown in the instances below:

1. a. How do I treat hand eczema?
- b. What can I use to relieve a sunburn?
- c. When is the best time to take your resting heart rate and why?
- d. Where can you find truthful answers about bone cancer?
- e. Why do we get blisters on our feet?

Both professional and consumer questions pose some difficulties for understanding due to typographical and grammatical errors. They also comprise instances that violate the general syntactic rules of written questions, e.g. some appear in declarative or imperative form rather than interrogative, or they are incomplete by the rules of any sentence type. Such instances include the following:

2. a. I wonder if this patient could have a rotator cuff thing? (professional)
- b. If im lactose intolerant. ....? (consumer)

Additionally, many of the linguistic phenomena that are in interrogative form are embedded within other sentences that are in other syntactic forms, as in the following:

3. a. This patient still has a cervical strain that is flared up. Muscle spasm? What to do next? I'll probably refer to a neurologist if still no better at the next visit (professional)
- b. I have mosquito bites on my feet and the scars from it aren't going away what can I do? (consumer)

In order to minimize some of the syntactic and pragmatics issues raised by such questions, we focused primarily on the shallow-level features (e.g. bag-of-words features) without deep language processing, which is discussed below. Additionally, the lexicon is of particular interest to the biomedical informatics community because of the challenges that medical terms pose for laypersons in terms of comprehension [37–41] and information retrieval [42,43].

### 3.3. Machine learning approach

The task is formulated as a binary classification problem aiming to separate consumer questions from professional questions

related to healthcare. We explored supervised machine-learning (ML) classifications with various algorithms using the freely available Weka Toolkit (<http://www.cs.waikato.ac.nz/ml/weka/>). To separate consumer questions from spoken clinical questions, we applied the classification framework on the Consumer–PointCare dataset where instances for two classes are from Consumer-I and PointCare (described in Section 3.2). Similarly, to separate consumer questions from online clinical question, we applied the classification framework on the Consumer–OnlinePractice dataset, where instances for two classes are from Consumer-II and OnlinePractice (described in Section 3.2), respectively.

#### 3.3.1. Learning features

**3.3.1.1. Bag-of-words (BOW) features.** As learning features for our ML approaches, we first explored bag-of-words, which relied on the lexical terms used in both sets of questions (referred as BOW features). To obtain BOW features from each question, we did the tokenization and filtered some noisy terms via regular expression, e.g. terms containing only symbols or starting with symbols.

**3.3.1.2. Statistical features.** Even though we used two relatively large collections of professional and consumer health questions (~10,000 each), the questions still represent only a small portion of questions that physicians and consumers could potentially pose to a question–answering system. Even between the two datasets we used there is much difference in the lexicon usage. In order to offset the data sparseness for a more robust system, we explored the use of features based on statistical aspects of language structure (referred as statistical features) listed as follows:

1. **Word length.** Healthcare professionals tend to use domain terms to express their information needs, and those technical terms are frequently longer than common words. We calculated the maximum, minimum and average letter counts of words in each question as features.
2. **Inverse Document Frequency (IDF).** Many domain terms in the professional questions are in rare usage compared with common words, and IDF is a good indicator of the corresponding word's rarity. We calculated each word's IDF value on the MEDLINE 2010 corpus that contains nearly 19 million records, and then used the maximum, minimum and average IDF value in each question as features.
3. **Question length.** Professional questions tend to contain more words and we counted the number of words in each question as a feature to capture the length difference between consumer questions and professional questions.

**3.3.1.3. Linguistic category features.** To improve the robustness of our approach, we extracted four linguistic categories as higher-level features (referred as linguistic category features) by manually

**Table 1**  
Top 20 words based on information gain value on the Consumer–PointCare dataset.

1	The	11	She
2	What	12	Of
3	Patient	13	You
4	My	14	A
5	Is	15	Help
6	With	16	Rid
7	Should	17	Dose
8	Woman	18	And
9	For	19	This
10	Her	20	In

examining the top bag-of-words features based on information gain based ranking.

Table 1 shows the top 20 words in the Consumer–PointCare dataset based on information gain. We observed that many of the terms that achieved a high information gain score were stopwords that are often omitted from NLP tasks including information retrieval due to their characterization as insignificant discriminators [44]. This finding is consistent with our earlier studies [21,30,45] in which we found stopwords were helpful for sentence classification and question classification tasks.

The finding that stopwords hold value for our classification has a practical benefit in that stopwords are typically members of the broad supercategory that is traditionally recognized in linguistics as closed-class words as opposed to open-class words [46]. Closed-class words belong to parts-of-speech sets that rarely, if ever, admit new members, including prepositions, pronouns, and interrogative words. The finite nature of these sets enabled us to extract the entire set as learning features based on the observation that some of the members of these sets had a high value in information gain. Based on these values, the second author of this study recognized four linguistic categories as having an especially high potential for our classification purposes, viz. interrogative words (e.g. *what*, *how*), personal pronouns (e.g. *I*, *my*, *her*, *she*, *you*), indefinite pronouns (e.g. *anyone*, *somebody*), and auxiliary verbs (e.g. *is*, *should*, *do*). As these categories are closed-class words, we extracted all the members from Wikipedia and automatically derived corresponding features indicating whether a question contains those linguistic category members. For open-class words such as nouns and verbs that do admit members freely, more sophisticated techniques are needed to group them in a higher level, which we will leave for our future work.

### 3.3.2. Evaluation metrics

We used recall, precision, and weighted F1-score as the evaluation metrics. Recall is the number of correctly classified sentences divided by the total number of sentences of that class; precision is the number of correctly classified sentences divided by the total number of sentences classified for the category; and the F1-score is the harmonic mean of recall and precision.

## 4. Results

In this section, we present the results of our supervised learning approach on the classification of consumer and professional questions. We experimented with different classification algorithms available in the Weka toolkit, and found that support vector machines (SVMs) with the Sequential Minimization Optimization (SMO) algorithm developed by Platt [47] worked best, and therefore, report only those results. We first evaluated our approach with different features on Consumer–PointCare and Consumer–OnlinePractice respectively using 10-fold cross-validation and we examined different characteristics between consumer questions and professional healthcare questions. The information gain based feature selection was then evaluated and finally we reported the classification performance applying the model trained using the Consumer–PointCare dataset on the Consumer–OnlinePractice dataset to show the robustness of our learning framework.

### 4.1. Performance comparison on two datasets with different features

We report 10-fold cross-validation results on two datasets (Consumer–PointCare and Consumer–OnlinePractice) separately, as shown in Table 2. We found the performance patterns with different feature settings are similar on two datasets. Bag-of-words (BOW) features perform best with the *F*-measure of 0.918 and

**Table 2**

Evaluation performance with different feature settings on two datasets.

Feature Settings	Precision	Recall	<i>F</i> -measure
<i>Results on Consumer–PointCare</i>			
Bag-of-words features (BOW)	0.918	0.918	0.918
Statistical features (SF)	0.825	0.824	0.824
BOW + SF	0.929	0.929	0.929
Linguistic category features (LCF)	0.821	0.818	0.818
LCF + SF	0.882	0.882	0.882
<i>Results on Consumer–OnlinePractice</i>			
Bag-of-words features (BOW)	0.945	0.945	0.945
Statistical features (SF)	0.876	0.875	0.875
BOW + SF	0.946	0.946	0.946
Linguistic category features (LCF)	0.757	0.755	0.755
LCF + SF	0.898	0.898	0.898

0.945 (row 3 and 10), compared with the 0.824/0.875 (row 4 and 11) using statistical features and 0.818/0.755 (row 6 and 13) using linguistic category features. Statistical features are shown to be very helpful in boosting the performance when combined with other features, achieving the highest *F*-measure of 0.929 on Consumer–PointCare and 0.946 on Consumer–OnlinePractice when combined with BOW features.

Overall the performance on Consumer–OnlinePractice is better than on Consumer–PointCare, except that using linguistic category features yielded better *F*-measure of 0.818 on Consumer–PointCare, compared with 0.755 on Consumer–OnlinePractice. Note that although the linguistic category features we proposed do not perform as well as other features, they provide more potential for better generalization ability than BOW features. We also tried stemmed BOW features and it degraded the performance, which is consistent with the finding in other natural language processing tasks [46,48].

To obtain a deep understanding of how each individual linguistic category and statistical feature contribute to distinguishing consumer questions from professional questions, we evaluated the performance on Consumer–PointCare dataset using each single corresponding feature respectively as shown in Table 3. The results on Consumer–OnlinePractice showed a similar pattern.

We can see most statistical features perform well on this task. The maximum word length yielded the best performance with an *F*-measure of 0.762, while the maximum IDF yielded the worst performance of 0.354. However, the average of IDF achieved a pretty good performance of 0.661, the third highest score following the second-best *F*-measure of 0.696 using question length. Of the four linguistic categories we proposed, auxiliary verbs performed the best with the *F*-measure of 0.726; indefinite pronouns performed the worst with the *F*-measure of 0.489. Interrogative words and personal pronouns performed similarly, yielding the *F*-measure of 0.673 and 0.657, respectively.

Fig. 1 shows the significant differences in the distribution of personal pronouns used in professional and consumer questions

**Table 3**

Analysis on the contribution of each statistical feature and linguistic category (*F*-measure).

Statistical features						
Word length			Question length		Inverse document frequency (IDF)	
Max	Avg	Min			Max	Avg
0.762	0.656	0.53	0.696		0.354	0.661
						0.534
Linguistic category features						
Interrogative words			Personal pronouns		Indefinite pronouns	
0.673			0.657		0.489	
					Auxiliary verbs	
					0.726	



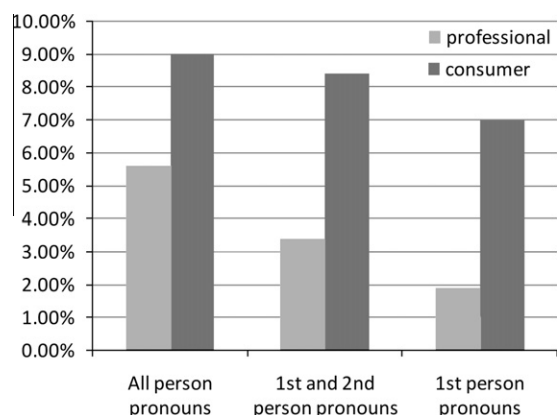


Fig. 1. Distributions of personal pronouns used in professional and consumer questions on Consumer-PointCare dataset.

(on Consumer-PointCare dataset), which further explains the effectiveness of using the personal pronoun linguistic category. As the figure shows, consumer questions incorporate a greater percentage of pronouns (5.6%) than professional questions (9.0%) overall, and there is a greater drop in usage among professional questions than consumer questions as the focus on pronouns is reduced from all pronouns to first- and second-person pronouns (3.4% versus 8.4%) and, finally, to only first-person pronouns (1.9% versus 7.0%).

Based on the analysis, we tried to remove the worst statistical feature (maximum word length) and the worst linguistic category (indefinite pronouns) in the corresponding feature settings of Table 2, but there was no improvement, showing that different features may complement each other through some interaction.

#### 4.2. Feature selection on BOW features

As previously described, bag-of-words (BOW) features are very useful for this classification; however, BOW features are computationally expensive to construct and they contribute to data sparseness problems, thereby posing a practical challenge on automatic question answering systems. In addition to the non-lexical statistical features and higher-level linguistic category features discussed earlier, we explored information gain based feature selection for dimension reduction on the BOW features, which are expected to reduce useless redundancies without noticeable information loss.

Fig. 2 shows the performance curves on two datasets with the different features selected by information gain. We can see that information gain based feature selection achieved the best F-measure of 0.926 on Consumer-PointCare and 0.946 on

Table 4

Classification performance using Consumer-PointCare as training data and Consumer-OnlinePractice as test data.

Feature settings	Precision	Recall	F-measure
Bag-of-words features (BOW)	0.88	0.879	0.879
Statistical features (SF)	0.858	0.843	0.841
BOW + SF	0.891	0.891	0.891
LCF + SF	0.86	0.856	0.856
BOW3000 + SF	0.891	0.891	0.891

Consumer-OnlinePractice compared with 0.918 and 0.945 using all the BOW features. In addition, using only 3000 top BOW features for both systems can achieve highly competitive performance and could potentially improve the system's robustness. When we combined with statistical features, using 3000 top BOW features based on information gain further improved the performance from 0.926 to 0.936 on Consumer-PointCare, and from 0.943 to 0.945 on Consumer-OnlinePractice.

#### 4.3. Performance when using Consumer-OnlinePractice as blind test data

In this section, we evaluated the robustness of our approaches with different settings. Specifically, we trained the learning model on the Consumer-PointCare dataset and used Consumer-OnlinePractice dataset as a blind test data for evaluation. The results are shown in Table 3, where "BOW + SF" for bag-of-words features combined with statistical features, "LCF + SF" for linguistic category features combined with statistical features, "BOW3000 + SF" for 3000 top BOW features based on information gain combined with statistical features.

We can see that using all BOW features is not as advantageous as the cross validation results on training data (Consumer-PointCare) in Table 2, achieving the F-measure of 0.879 compared with 0.918 (row 3 in Table 2), while statistical features as expected were more robust on new test data and obtained the F-measure of 0.841 compared with 0.824 (row 4 in Table 2). It also shows that the proposed linguistic category features (LCF) and the dimension reduction of BOW features (BOW3000) based on information gain score both help gain better performance on the test data when combining with the statistical features, with the F-measure increasing from 0.841 to 0.856 and the best 0.891 respectively. In addition, we found keep the top 3000 BOW features did not lose any useful information, yielding the same best performance as using all the BOW features (compare rows 4 and 6 in Table 4).

## 5. Discussion

This study explored the possibility of using a classifier to distinguish between questions asked by healthcare professionals and those asked by consumers. Although the semantics of medical questions play an important role for related tasks, the study focused on the language usage difference between consumers and healthcare professionals, where traditional semantics would not matter as much as other question classification tasks. The results show that our supervised learning framework based on inexpensive bag-of-words features and statistical features obtained satisfying performance for this classification task.

To account for the diversity between point-of-care and online physician practice questions, the study used two datasets Consumer-PointCare and Consumer-OnlinePractice in our study. From the results in Table 2, we can see that the different language facets in two different clinical settings influenced the accuracy of distinguishing consumer questions from them, but similar patterns with different feature settings prove that they share internal

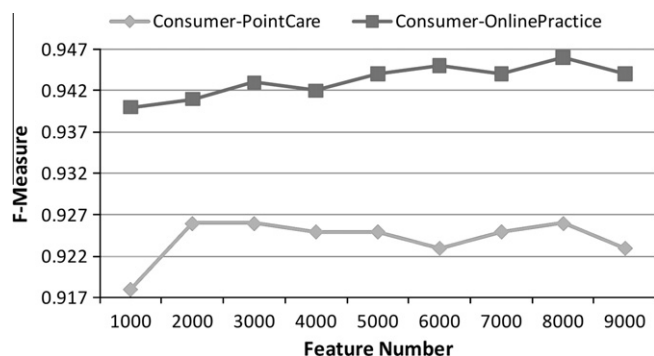


Fig. 2. Performance curve with different number of features selected based on information gain (based on F-measure).

characterizations, which makes it quite feasible to develop an effective and unified learning framework compatible with both point-of-care and online physician practice questions, even other new professional questions generated under different scenarios.

In Table 4 for the training model on the Consumer–PointCare dataset and testing on the Consumer–OnlinePractice dataset, the performance is slightly worse than the cross-validation results on the training data (Consumer–PointCare) shown in Table 2, but the nearly 0.9 of the *F*-measure still shows the strong robustness of the framework proposed for this task. In other words, although there are some differences in various aspects of language usage between point-of-care and online physician practice questions, this result further validates that the difference in those two type of professional questions will not overwhelm their inherent difference from consumer questions. This finding lays a solid foundation towards further development in automated consumer question answering.

Among the seven statistical-based features we used, word length was shown to be a significant indicator of the identity of the question askers. We found that professionals are more likely than consumers to include long scientific words in their questions. This finding has been previously observed in the literature on differences between the language of healthcare professionals and consumers; for instance, it has been addressed by the U.S. Food and Drug Administration has proposed simpler language on medication labels by replacing words such as “pulmonary” with “lung”, “assistance” with “help” or “aid”, and “medication” with “drug” [41].

As shown in Table 1, the number of words in the question also played an important role in question classification. Among the individual features analyzed on the Consumer–PointCare dataset in Table 3, Maximum Word Length achieved the best performance with the *F*-measure of 0.762, suggesting that the appearance of long words in a question is indicative of the type of user that asked the question. We found in the Consumer–PointCare dataset the maximum length of individual words differed considerably between two types of questions: the longest word in the professional questions, for instance, is 27 letters long (*esphogagogastroduodenoscopy*), and the professional question set contains numerous words comprising 20 letters or more, including *electroencephalographic*, *pseudothrombophlebitis*, and *dehydroepiandrosterone*. The longest word in the consumer question, on the other hand, is a single instance of the misspelled, 20-letter word *herpaphonacyphilaid*.

The overall distributions of word length on the two types of questions also differed. While we found that in the professional questions of the Consumer–PointCare dataset, 24.5% of the words are 7 or more letters and only 15.3% of the words for the consumer questions are 7 or more letters; 7.6% of the words in the professional questions are 10 or more letters compared with only 2.6% of consumer questions with the same length. From these numbers, we can see that the greater the length of the word, the more likely it is to appear in a question asked by a healthcare professional rather than a consumer.

We assumed that highly educated and specialized healthcare professionals use more domain terms than consumers, making inverse document frequency (IDF) a useful feature for classification; however, our analysis on each individual feature in Table 3 showed mixed results. While the average IDF works quite well, the maximum IDF achieved the worst *F*-measure among all the other statistical features, suggesting that assumption to be unwarranted, for several possible reasons. First, the misspellings and Internet abbreviations that have appeared in recent years, such as “plz” and “lol,” and appear in consumer questions are difficult to account for using such a model. Furthermore, in addition to such forms, consumer questions might include infrequent dialectal variants for diseases, treatments, and conditions that are unlikely to surface in general

medical language. Finally, from a methodological standpoint, we calculated IDF values from the MEDLINE corpus; however, basing IDF values on the distribution of words in a more general corpus or from a variety of corpora might be a better strategy for making the most of IDF value for such classification.

We observed that bag-of-words (BOW) features contain many redundancies for the classification task. With the increase of number of BOW features as shown in Fig. 2, the performance has only marginal improvements or even degradations, which is why information gain based dimension reduction allows the system to use fewer BOW features without information loss, which is especially useful to making the system more compatible and robust as shown in Table 4.

The proposed linguistic category features achieved promising results on our classification task. Their effectiveness was demonstrated in both cross-validation results (Table 2) and blind test results (Table 4), boosting the performance when combined with statistical features. Although the linguistic category features are not obtaining the best performance in our current study, it provides a potential way in the future to examine other linguistic categories (specific open-class sets, e.g. nominalizations and words of Latin and Greek origin) or topic-related clusters that could further benefit from this classification task. We especially analyzed the distributions of one linguistic category (personal pronouns) as shown in Fig. 1, suggesting that differences in subjectivity and objectivity play important roles in distinguishing between the language of professional questions and that of consumer questions. In other words, pronoun usage in the professional questions reflects the objective orientation of clinicians who generally have questions about the healthcare of their patients and infrequently refer to themselves in medical questions. Consumer questions, on the other hand, generally concern he askers of question because they are experiencing the problem directly and their pronoun usage reflects their subjectivity. Although these observations are perhaps intuitively obvious, our study shows they have practical applications for classification.

Although our systems perform quite well, there are several limitations in the current study. With respect to the use of IDF values as a means for measuring the value of words based on their rarity, we used MEDLINE as our reference corpus with limited success. Future work on this aspect of language use might be better served by relying on more general English corpora, such as the Brown Corpus, the Frown Corpus or the Wall Street Journal Corpus, or by using a combination of medical and general corpora. Additionally, the word count of a sentence was shown to be useful for this task; however, future work should explore the use of more sophisticated measurements of sentential complexity, such as deep linguistic analysis (e.g. automatic parsing) of questions might also aid in this kind of classification and will be addressed in future work. Finally, only Yahoo prompts were tested for consumer questions; future research should include all the language provided by the asker in Yahoo answers for a given linguistic event.

## 6. Conclusion

We evaluated a supervised learning framework for classifying consumer questions from professional questions, in which we explored bag-of-words features as well as statistical features. The results of our work suggest that automating the classification of questions into professional and consumer categories is feasible. The proposed approach performed well in separating consumer questions from two types of professional questions (point-of-care and online physician practice), and the competitive performance was generalized when training a model on Consumer–PointCare and testing on Consumer–OnlinePractice, showing the robustness

of our learning framework on this specific task. The proposed linguistic category features and dimension reduction on bag-of-words features were shown to enhance the system's robustness. In addition, several differences between the questions of health-care professionals and health consumers were analyzed, such as word length and personal pronoun usage.

Our future work will further enhance the classification performance by exploring more helpful features, such as using IDF calculated with a more general corpus, linguistic open classes or topic related word clusters, and syntactic parsing. More extensive evaluations will be conducted in the future with user central evaluation. We will investigate a systematic way to incorporate the classification framework proposed in this paper into the ASKHermes system for a preliminary evaluation on automated consumer question answering systems.

## Acknowledgments

The authors would like to thank Yong-Gang Cao for technical support and Shashank Agarwal and Betsy Barry for helpful comments.

## References

- [1] Anderson JG. Consumers of e-health: patterns of use and barriers. *Soc Sci Comput Rev* 2004;22:242–8.
- [2] Elkin N. How America searches: health and wellness. *iCrossing* 2008;1–17.
- [3] Weaver III JB, Mays D, Lindner G, Eroglu D, Fridinger F, Bernhardt JM. Profiling characteristics of internet medical information users. *J Am Med Inform Assoc* 2009;16:714–22.
- [4] Tu HT, Cohen GR. Striking jump in consumers seeking health care information. *Track Rep* 2008;1–8.
- [5] Landro L. Social networking comes to health care. *Wall Street J* 2006;27.
- [6] Agichtein E, Castillo C, Donato D, Gionis A, Mishne G, Agichtein E, et al. Finding high-quality content in social media with an application to community-based question answering. In: *Proceedings of WSDM*; 2008.
- [7] Yu H, Cao YG. Automatically extracting information needs from ad hoc clinical questions. In: *AMIA annual symposium proceedings 2008*; 2008. p. 96.
- [8] Cao YG, Ely J, Antieau L, Yu H. Evaluation of the clinical question answering presentation. In: *BioNLP*; 2009.
- [9] Liu F, Tur G, Hakkani-Tür D, Yu H. Towards spoken clinical question answering: evaluating and adapting automatic speech recognition systems for spoken clinical questions. *J Am Med Inform Assoc* 2011;18:625–30.
- [10] Dorr DA, Tran H, Gorman P, Wilcox AB. Information needs of nurse care managers. In: *AMIA annu symp proc*; 2006. p. 913.
- [11] Katz MG, Jacobson TA, Veledar E, Kripalani S. Patient literacy and question-asking behavior during the medical encounter: a mixed-methods analysis. *J Gen Intern Med* 2007;22:782–6.
- [12] Graber MA, Randles BD, Ely JW, Monahan J. Answering clinical questions in the ED. *Am J Emerg Med* 2008;26:144–7.
- [13] Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, et al. Analysis of questions asked by family doctors regarding patient care. *BMJ* 1999;319:358–61.
- [14] Roter DL. Patient participation in the patient–provider interaction: the effects of patient question asking on the quality of interaction, satisfaction and compliance. *Health Educ Behav* 1977;5:281.
- [15] Rudd RE, Moeykens BA, Colton TC. Health and literacy. A review of medical and public health literature. *Annual review of adult learning and literacy*. New York: Jossey-Bass; 1999. <<http://www.cete.org/acve/docs/pab00016.pdf>>.
- [16] Bader JL, Theofanos MF. Searching for cancer information on the internet: analyzing natural language search queries. *J Med Internet Res* 2003;5.
- [17] Mendonça EA, Kaufman D, Johnson SB. Answering information needs in workflow.
- [18] Joachims T. Learning to classify text using support vector machines: methods, theory, and algorithms. *Comput Linguis* 2002;29:656–64.
- [19] Suzuki J, Taira H, Sasaki Y, Maeda E. Question classification using HDAG kernel. In: *Proceedings of the ACL 2003 workshop on multilingual summarization and question answering*, vol. 12; 2003. p. 61–8.
- [20] Hermjakob U. Parsing and question classification for question answering. In: *ACL workshop on open-domain question answering*; 2001.
- [21] Yu H, Sable C, Zhu HR. classifying medical questions based on an evidence taxonomy. In: *Proceedings of the AAAI 2005 workshop on question answering in restricted domains*; 2005.
- [22] Zhang D, Lee WS. Question classification using support vector machines. In: *The 26th annual international ACM SIGIR conference*; 2003. p. 26–32.
- [23] Li X, Doth D. Learning question classifiers. In: *COLING'02*; 2002.
- [24] Hacıoglu K, Ward W. Question classification with support vector machines and error correcting codes. In: *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology: companion volume of the proceedings of HLT-NAACL 2003-short papers*, vol. 2; 2003. p. 28–30.
- [25] Small S, Liu T, Shimizu N, Strzalkowski T. HITQA: an interactive question answering system a preliminary report. In: *Proceedings of the ACL 2003 workshop on multilingual summarization and question answering*, vol. 12; 2003. p. 46–53.
- [26] Small S, Strzalkowski T. HITQA: a data driven approach to interactive analytical question answering. In: *Proceedings of HLT-NAACL 2004: short papers on XX*; 2004. p. 53–6.
- [27] Ku L-wei, Liang Y-ting, Chen H-hsi. Question analysis and answer passage retrieval for opinion question answering systems.
- [28] Li B, Liu Y, Agichtein E. CoCQA: co-training over questions and answers with an application to predicting question subjectivity orientation. In: *Proceedings of the conference on empirical methods in natural language processing*; 2008. p. 937–46.
- [29] Li B, Liu Y, Ram A, Garcia EV, Agichtein E. Exploring question subjectivity prediction in community QA. In: *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*; 2008. p. 735–6.
- [30] Yu H, Sable C. Being Erlang Shen: identifying answerable questions. In: *Proceedings of the nineteenth international joint conference on artificial intelligence on knowledge and reasoning for answering questions*; 2005.
- [31] Biber D, Conrad S, Reppen R. *Corpus linguistics: investigating language structure and use*. Cambridge Univ Pr.; 1998.
- [32] Biber D. Variation across speech and writing. Cambridge Univ Pr.; 1991.
- [33] Redish JC, Selzer J. The place of readability formulas in technical communication. *Tech Commun* 1985;32:46–52.
- [34] Lin J, Demner-Fushman D. “Bag of Words” is not enough for Strength of evidence classification. In: *AMIA annual symposium proceedings 2005*; 2005. p. 1031.
- [35] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In: *Proceedings of the fourteenth international conference (ICML'97)*; 1997.
- [36] Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, et al. A taxonomy of generic clinical questions: classification study. *BMJ* 2000;321:429–32.
- [37] Hartley J. Clarifying the abstracts of systematic literature reviews. *Bull Med Lib Assoc* 2000;88:332.
- [38] McCray AT, Ide NC, Loane RR, Tse T. Strategies for supporting consumer health information seeking. In: *Medinfo 2004: proceedings of the 11th world conference on medical informatics*, San Francisco, September 7–11, 2004; 2004. p. 1152.
- [39] McCray AT. Promoting health literacy. *J Am Med Inform Assoc* 2005;12:152–63.
- [40] Weeks WB, Wallace AE. Readability of British and American medical prose at the start of the 21st century. *BMJ* 2002;325:1451.
- [41] Farley D. Label Literacy for OTC Drugs. FDA consumer; 1997. p. 31.
- [42] Zeng QT, Tse T. Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc* 2006;13:24.
- [43] Zeng QT, Crowell J, Plovnick RM, Kim E, Ngo L, Dibble E. Assisting consumer health information retrieval with query recommendations. *J Am Med Inform Assoc* 2006;13:80–90.
- [44] Yates RB, Neto BR. *Modern information retrieval*. New York: ACM Press; 1999.
- [45] Agarwal S, Yu H. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. In: *Proceedings of the AMIA summit on translational bioinformatics*; 2009.
- [46] Jurafsky D, Martin JH. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall PTR; 2000.
- [47] Platt JC. Using analytic QP and sparseness to speed training of support vector machines. *Adv Neural Inform Process Syst* 1999;9:557–63.
- [48] Hull DA. Stemming algorithms: a case study for detailed evaluation. *J Am Soc Inform Sci* 1996;47:70–84.